用户画像：方法论与工程化解决方案（从技术、产品、运营3个维度详尽阐述从0到1搭建用户画像系统的技术栈和方法论）

作者：赵宏田 AiBooKs.cc

用户画像：方法论与工程化解决方案

赵宏田　著

ISBN：978-7-111-63564-2

本书纸版由机械工业出版社于2019年出版，电子版由华章分社（北京华章图文信息有限公司，北京奥维博世图书发行有限公司）全球范围内制作与发行。

客服热线：+ 86-10-68995265

客服信箱：service@bbbvip.com

官方网址：www.hzmedia.com.cn

新浪微博 @华章数媒

微信公众号 华章电子书（微信号：hzebook）

目录

前言

为什么写这本书

我曾在知乎“数据智能”专栏下面不定期连载关于用户画像的文章，也曾在知乎开设过几期live直播，还曾在天善智能等网课平台开设过系列网课“用户画像解决方案”。在和同行业中对画像感兴趣的朋友们交流时，我发现大家虽然来自地产、烟草、零售、互联网等不同行业，但所在公司对用户画像领域都有建设需求，而且大家对于指标体系、标签作业效率（ETL）、标签监控、实时计算、画像产品化、业务应用场景和应用方式等方面都有进一步了解的兴趣。所以我想对这些年做用户画像的经验、踩过的“坑”进行梳理总结，为数据开发、数据分析、运营、用户研究等岗位的工作人员提供一些参考。

在写这份解决方案的一个个夜晚，我有时会想，科技和时代都在飞速发展，如果有一天我不做这一行了，该拿什么来回忆那些年奋斗的时光呢？2019年，我第3次从0到1开始搭建用户画像系统，从离线标签开发、用户数据分析、ETL调度、流式计算开发，到打通数据服务层、应用画像数据服务业务方、获得业务增长的反馈，这一路走过来，过程是痛苦的，收获是丰富的。奋斗的日子固然多彩，回望一步步走过的路，谨以此书向那些不舍昼夜奔腾向前的日子致敬。

本书特色

开始做用户画像的时候我也不知道从何处下手，市面上介绍Hive、Spark、HBase、MySQL、数据仓库等大数据相关技术的书籍很多，但是介绍用户画像搭建开发的书籍很少，甚至没有。在没有相关项目经验的情况下，我不知道如何把这些大数据组件统筹起来搭建用户画像系统。直到这两年，我才一边开发画像系统，一边总结梳理，最终编纂成本书。

本书借助数据仓库实现一套用户画像系统的方案。从实际工程案例出发，结合多业务场景，内容涵盖开发离线批处理计算的标签及流式计算标签，为读者的分析、开发、搭建用户画像系统，并借助该用户画像系统为运营人员制定运营用户的策略提供端到端的解决方案。

一套好的解决方案需要包括以下几个层面。

1）架构层：在画像系统的架构层，本书首先介绍了画像数据仓库的架构，进一步介绍了数据存储的技术选型，在什么场景下使用Hive、MySQL、HBase、Elasticsearch等工具存储数据，用户标签开发、人群计算开发等相应数据开发层面的内容，以及整个项目的开发流程和各阶段的关键产出。

2）流量层：介绍整个方案是如何运作起来的。本书主要涉及画像系统的作业流程调度、数据仓库和各业务系统的打通。

3）业务层：包括系统的前后端交互以及如何把这套系统应用在业务服务层面。本书通过用户画像产品化介绍了产品端和画像系统的“代码”层面是如何进行交互操作的。

4）方案价值：包括系统上线后如何服务于各业务场景产生业务价值以及有待进一步完善的地方。

以上几个层面的内容构成了一套完整的用户画像解决方案，这也是本书各章节覆盖的全部模块。

数据的最终目的是走出数据仓库，应用到业务系统和营销系统中来驱动营收增长。

我在学习数据仓库的时候学过Kimball的《数据仓库工具箱》，其中关于数据仓库的34个子系统的介绍对我影响很大，其对于如何解决特定问题并形成结构化思维有着系统的方法论与解决方案。虽然面对具体问题的处理方式是灵活且丰富多样的，但是固定的结构化思维有利于快速找到突破口，形成良好的开端。

本书可以帮助读者在用户画像领域形成一种体系化思维，在面对一个具体项目时不会无从下手。如何建立标签指标体系？指标体系中包含哪些标签？如何设计存储画像标签的表结构？如何开发标签？画像系统中涉及哪些数据存储工具？如何打通标签数据到服务层？如何对画像系统进行监控？如何对整个画像系统进行调度？如何使画像系统服务于业务场景来驱动增长？这些都是画像系统的子模块。

主要章节及内容

本书共9章，各章具体内容如下：

第1章：主要讲用户画像的基础知识，包括搭建用户画像系统需要覆盖的模块，开发阶段流程，各阶段的关键产出，以及数据仓库架构、表结构的设计等内容。阅读本章可以帮助读者形成构建用户画像的一个整体化思想。

第2章：结合业务设定指标体系，本章针对案例背景，从常用的用户属性、行为、消费、风险控制这4个维度设定指标体系。本章提供的标签可涵盖大部分刻画用户画像的应用场景，对于具体应用点，读者可根据公司业务特性进行针对性的补充。

第3章：讲解了标签相关数据的存储，包括Hive存储、MySQL存储、HBase存储和Elasticsearch存储。不同的存储方式适用于不同的场景和业务需要。

第4章：也是本书的重点章节，书中介绍的标签数据及相关脚本的开发是用户画像构建工作的重点。本章讲解了对常见的统计类、规则类、挖掘类、流式计算类标签以及用户特征库等与用户相关的数据的开发，还进一步介绍了如何计算人群数据、打通数据到服务层通路的开发。通过GraphX图计算用户2度关系熟人的案例介绍了如何深度挖掘用户间的关联关系。本章对每一小节都进行了详细的讲解，并附有配套的代码计算过程。

第5章：讲解了开发过程中常见的数据倾斜调优、对小文件的读取、缓存中间数据、开发中间表等调优工作。

第6章：讲解了如何使用开源ETL工具Airflow实现画像系统相关任务的工程化上线调度，以及对数据的监控预警和调度异常的排查。

第7章：画像产品化是数据从数据仓库走向业务服务的重要环节，画像产品化可便于业务人员使用工具来分析用户，将业务上定义的用户群应用到各业务系统中提供服务。本章为数据产品人员、业务人员提供了解决方案的思路。

第8章：介绍了用户画像的应用场景，包括经营分析、精准营销、个性化推荐等应用方向，方便业务人员、产品经理、数据分析师更好地了解用户、触达用户。

第9章：通过场景化介绍用户画像实际应用的8个案例，清楚地展现了用户画像作为一种分析、触达用户的工具在实际业务上的应用方式和应用流程。

主要读者对象

·产品经理：由于岗位性质对技术不是特别熟悉，可重点关注第1、2、7、8、9章的内容。

·数据分析师：可以从多个维度对用户及用户群进行分析，可重点关注第1、2、3、7、8、9章的内容。

·运营人员：可重点关注第2、8、9章的内容，了解画像系统涉及的指标体系、应用场景及应用策略。

·数据开发人员：本书主要站在数据开发人员的角度对整个画像系统进行系统化介绍。数据开发人员可完整阅读本书各章的内容。

·市场人员：借助画像系统了解用户群体的特征以及运营用户群的策略方法，可重点关注第2、8、9章的内容。

勘误和支持

由于水平有限，书中难免会存在疏漏之处，恳请读者批评指正。为此，读者可通过邮箱（892798505@qq.com）或微信（administer00001）反馈有关问题，我将尽全力为读者进行解答。

致谢

感谢父母对我一路成长的支持。感谢机械工业出版社华章公司的杨福川老师和李艺老师，这是我第二次与两位老师合作，每次合作与沟通总是那么愉快；感谢为本书写推荐的朋友们，你们的专业建议让本书更加精彩。最后，感谢过去一年中自己的每一分投入，不断积累，将大数据在用户画像领域的工程化实现和应用方案编纂成书。

第1章　用户画像基础

1.1　用户画像是什么

在互联网步入大数据时代后，用户行为给企业的产品和服务带来了一系列的改变和重塑，其中最大的变化在于，用户的一切行为在企业面前是可“追溯”“分析”的。企业内保存了大量的原始数据和各种业务数据，这是企业经营活动的真实记录，如何更加有效地利用这些数据进行分析和评估，成为企业基于更大数据量背景的问题所在。随着大数据技术的深入研究与应用，企业的关注点日益聚焦在如何利用大数据来为精细化运营和精准营销服务，而要做精细化运营，首先要建立本企业的用户画像。

1.1.1 画像简介

用户画像，即用户信息标签化，通过收集用户的社会属性、消费习惯、偏好特征等各个维度的数据，进而对用户或者产品特征属性进行刻画，并对这些特征进行分析、统计，挖掘潜在价值信息，从而抽象出用户的信息全貌，如图1-1所示。用户画像可看作企业应用大数据的根基，是定向广告投放与个性化推荐的前置条件，为数据驱动运营奠定了基础。由此看来，如何从海量数据中挖掘出有价值的信息越发重要。

图1-1　某用户标签化

大数据已经兴起多年，其对于互联网公司的应用来说已经如水、电、空气对于人们的生活一样，成为不可或缺的重要组成部分。从基础设施建设到应用层面，主要有数据平台搭建及运维管理、数据仓库开发、上层应用的统计分析、报表生成及可视化、用户画像建模、个性化推荐与精准营销等应用方向。

很多公司在大数据基础建设上投入很多，也做了不少报表，但业务部门觉得大数据和传统报表没什么区别，也没能体会大数据对业务有什么帮助和价值，究其原因，其实是“数据静止在数据仓库，是死的”。

而用户画像可以帮助大数据“走出”数据仓库，针对用户进行个性化推荐、精准营销、个性化服务等多样化服务，是大数据落地应用的一个重要方向。数据应用体系的层级划分如图1-2所示。

图1-2　数据应用体系的层级划分

1.1.2　标签类型

用户画像建模其实就是对用户“打标签”，从对用户打标签的方式来看，一般分为3种类型（如图1-3所示）：①统计类标签；②规则类标签；③机器学习挖掘类标签。

图1-3　标签类型

下面我们介绍这3种类型的标签的区别：

1.统计类标签

这类标签是最为基础也最为常见的标签类型，例如，对于某个用户来说，其性别、年龄、城市、星座、近7日活跃时长、近7日活跃天数、近7日活跃次数等字段可以从用户注册数据、用户访问、消费数据中统计得出。该类标签构成了用户画像的基础。

2.规则类标签

该类标签基于用户行为及确定的规则产生。例如，对平台上“消费活跃”用户这一口径的定义为“近30天交易次数≥2”。在实际开发画像的过程中，由于运营人员对业务更为熟悉，而数据人员对数据的结构、分布、特征更为熟悉，因此规则类标签的规则由运营人员和数据人员共同协商确定；

3.机器学习挖掘类标签

该类标签通过机器学习挖掘产生，用于对用户的某些属性或某些行为进行预测判断。例如，根据一个用户的行为习惯判断该用户是男性还是女性、根据一个用户的消费习惯判断其对某商品的偏好程度。该类标签需要通过算法挖掘产生。

在项目工程实践中，一般统计类和规则类的标签即可以满足应用需求，在开发中占有较大比例。机器学习挖掘类标签多用于预测场景，如判断用户性别、用户购买商品偏好、用户流失意向等。一般地，机器学习标签开发周期较长，开发成本较高，因此其开发所占比例较小。

1.2 数据架构

在整个工程化方案中，系统依赖的基础设施包括Spark、Hive、HBase、Airflow、MySQL、Redis、Elasticsearch。除去基础设施外，系统主体还包括Spark Streaming、ETL、产品端3个重要组成部分。图1-4所示是用户画像数仓架构图，下面对其进行详细介绍。

图1-4 用户画像数仓架构

图1-4下方虚线框中为常见的数据仓库ETL加工流程，也就是将每日的业务数据、日志数据、埋点数据等经过ETL过程，加工到数据仓库对应的ODS层、DW层、DM层中。

中间的虚线框即为用户画像建模的主要环节，用户画像不是产生数据的源头，而是对基于数据仓库ODS层、DW层、DM层中与用户相关数据的二次建模加工。在ETL过程中将用户标签计算结果写入Hive，由于不同数据库有不同的应用场景，后续需要进一步将数据同步到MySQL、HBase、Elasticsearch等数据库中。

·Hive：存储用户标签计算结果、用户人群计算结果、用户特征库计算结果。

·MySQL：存储标签元数据，监控相关数据，导出到业务系统的数据。

·HBase：存储线上接口实时调用类数据。

·Elasticserch：支持海量数据的实时查询分析，用于存储用户人群计算、用户群透视分析所需的用户标签数据（由于用户人群计算、用户群透视分析的条件转化成的SQL语句多条件嵌套较为复杂，使用Impala执行也需花费大量时间）。

用户标签数据在Hive中加工完成后，部分标签通过Sqoop同步到MySQL数据库，提供用于BI报表展示的数据、多维透视分析数据、圈人服务数据；另一部分标签同步到HBase数据库用于产品的线上个性化推荐。

1.3 主要覆盖模块

搭建一套用户画像方案整体来说需要考虑8个模块的建设，如图1-5所示。

·用户画像基础：需要了解、明确用户画像是什么，包含哪些模块，数据仓库架构是什么样子，开发流程，表结构设计，ETL设计等。这些都是框架，大方向的规划，只有明确了方向后续才能做好项目的排期和人员投入预算。这对于评估每个开发阶段重要指标和关键产出非常重要，重点可看1.4节。

·数据指标体系：根据业务线梳理，包括用户属性、用户行为、用户消费、风险控制等维度的指标体系。

·标签数据存储：标签相关数据可存储在Hive、MySQL、HBase、Elasticsearch等数据库中，不同存储方式适用于不同的应用场景。

·标签数据开发：用户画像工程化的重点模块，包含统计类、规则类、挖掘类、流式计算类标签的开发，以及人群计算功能的开发，打通画像数据和各业务系统之间的通路，提供接口服务等开发内容。

图1-5　用户画像主要覆盖模块

·开发性能调优：标签加工、人群计算等脚本上线调度后，为了缩短调度时间、保障数据的稳定性等，需要对开发的脚本进行迭代重构、调优。

·作业流程调度：标签加工、人群计算、同步数据到业务系统、数据监控预警等脚本开发完成后，需要调度工具把整套流程调度起来。本书讲解了Airflow这款开源ETL工具在调度画像相关任务脚本上的应用。

·用户画像产品化：为了能让用户数据更好地服务于业务方，需要以产品化的形态应用在业务上。产品化的模块主要包括标签视图、用户标签查询、用户分群、透视分析等。

·用户画像应用：画像的应用场景包括用户特征分析、短信、邮件、站内信、Push消息的精准推送、客服针对用户的不同话术、针对高价值用户的极速退货退款等VIP服务应用。

本书内容安排也分别围绕这8个模块的内容来展开。方便读者更清楚地了解用户画像是如何从0到1搭建起来并提供服务、驱动用户和实现营收增长的。